Sales Analysis & Forecasting for a Retail Company

Mahishi Rajaguru

June 26, 2025

1 Introduction

This project analyzes historical sales data from a retail company, applying modern data analytics and machine learning techniques for insight generation and forecasting. The objective is to demonstrate data wrangling, exploratory data analysis (EDA), visualization, and time series forecasting using both ARIMA and Prophet models.

2 Data Overview

- **Dataset:** Superstore US Retail Transactions
- **Rows:** 9,994 **Columns:** 21
- **Main fields:** Order/Ship dates, Product Category, Region, Sales, Profit, Discount, etc.
- **Time Range:** 2014–2017
- **Missing Data:** None detected

3 Exploratory Data Analysis (EDA)

3.1 Total Sales and Profits

- **Total Sales:** $2,297,201.07
- **Total Profit:** $286,397.79
- **Number of Orders:** 5,009

3.2 Sales by Category and Region

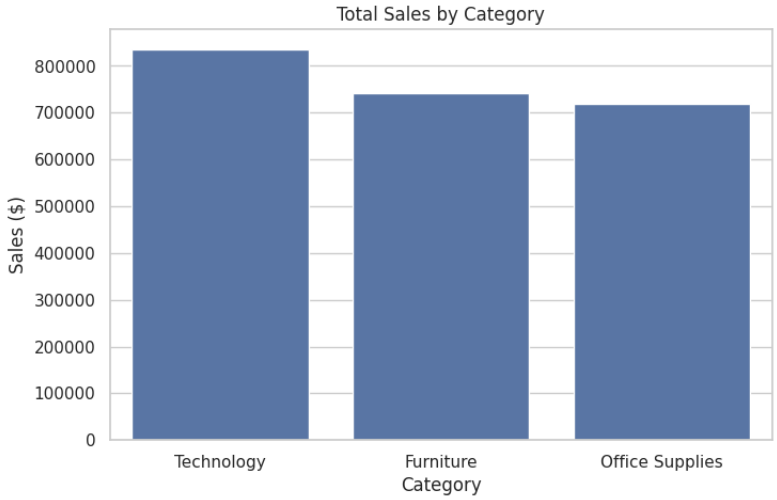

Figure 1: Total Sales by Category

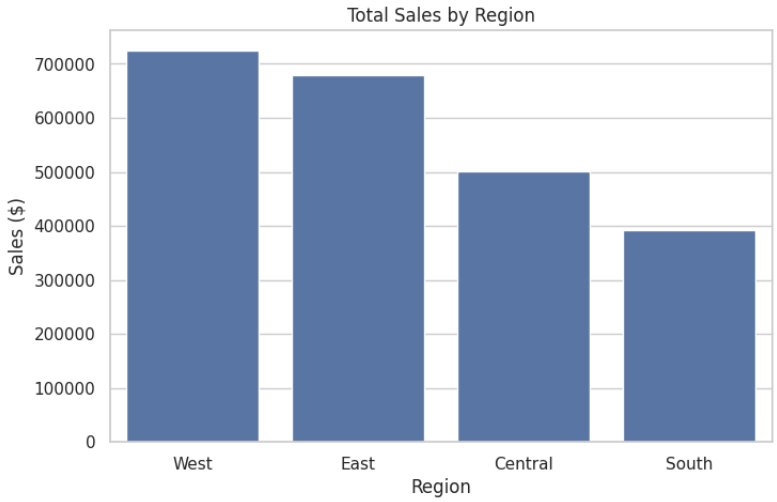

Figure 2: Total Sales by Region

3.3 Monthly Sales Trends

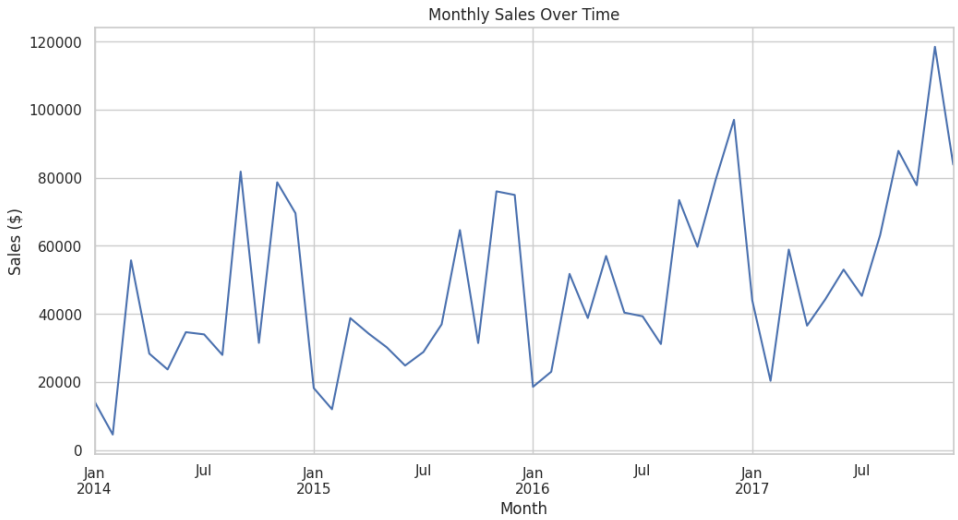

Figure 3: Monthly Sales Over Time

3.4 Top Products and Profitability

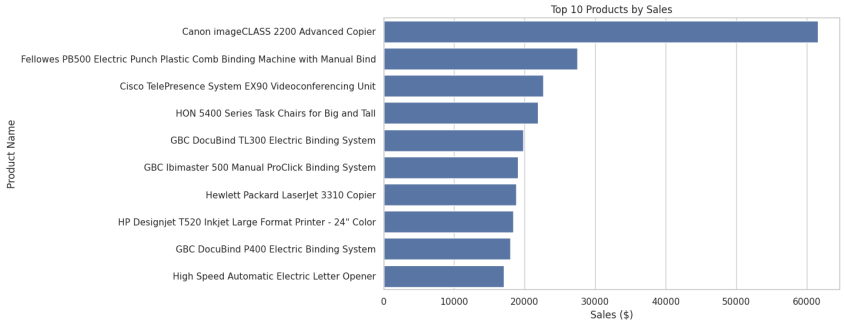

Figure 4: Top 10 Products by Sales

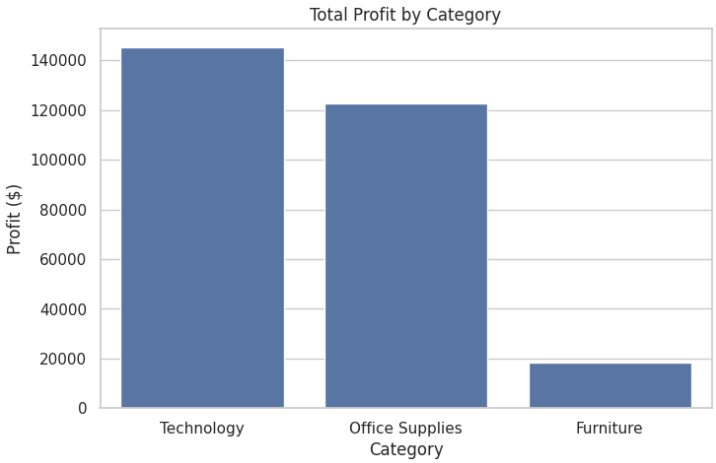

Figure 5: Total Profit by Category

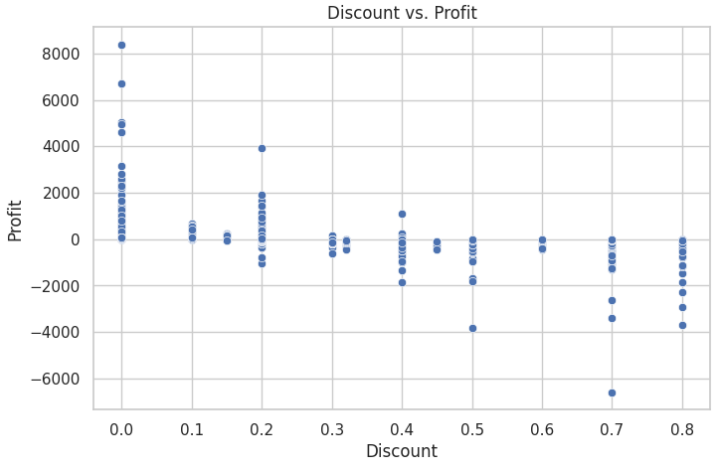

Figure 6: Discount vs. Profit Scatterplot

There is a negative relationship between discount rates and profit. Excessive discounting often results in reduced profitability, underscoring the need for careful discount management.

4 Time Series Forecasting

4.1 ARIMA Model

ARIMA (AutoRegressive Integrated Moving Average) is a classical approach for forecasting time series data. The ARIMA(1,1,1) model was used to capture the trend and predict future sales.

Theory

ARIMA stands for *AutoRegressive Integrated Moving Average*. It is used for analyzing and forecasting time series data.

The general ARIMA(p,d,q) equation is:

$$y_t = c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where:

- y_t: value at time t
- ϕ_i: coefficients for autoregressive terms
- θ_j: coefficients for moving average terms
- ε_t: error term (white noise)

4.2 ACF and PACF

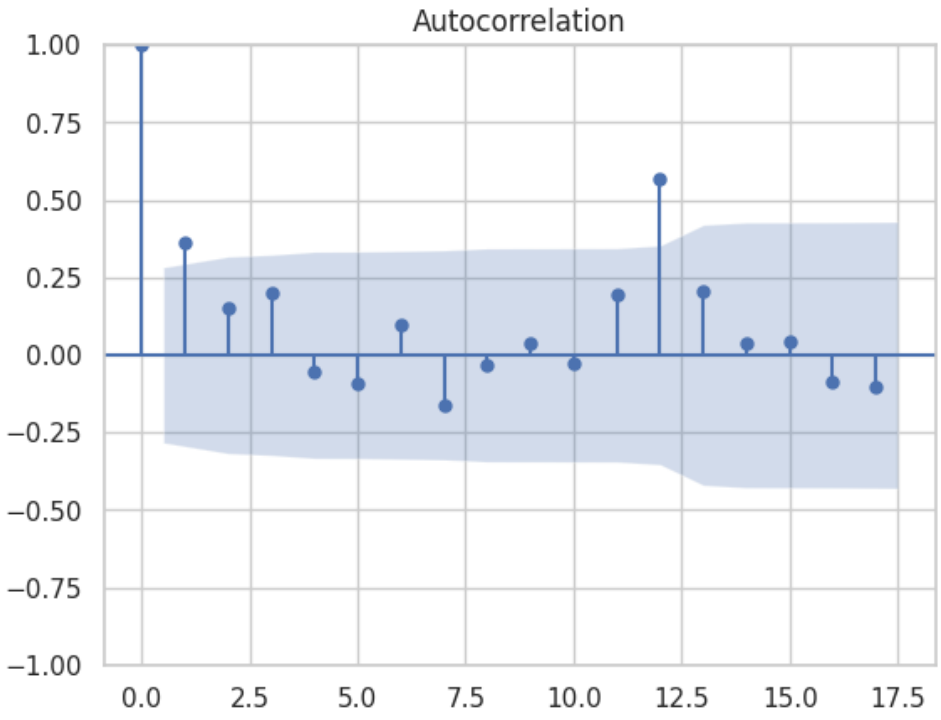

Figure 7: ACF sales

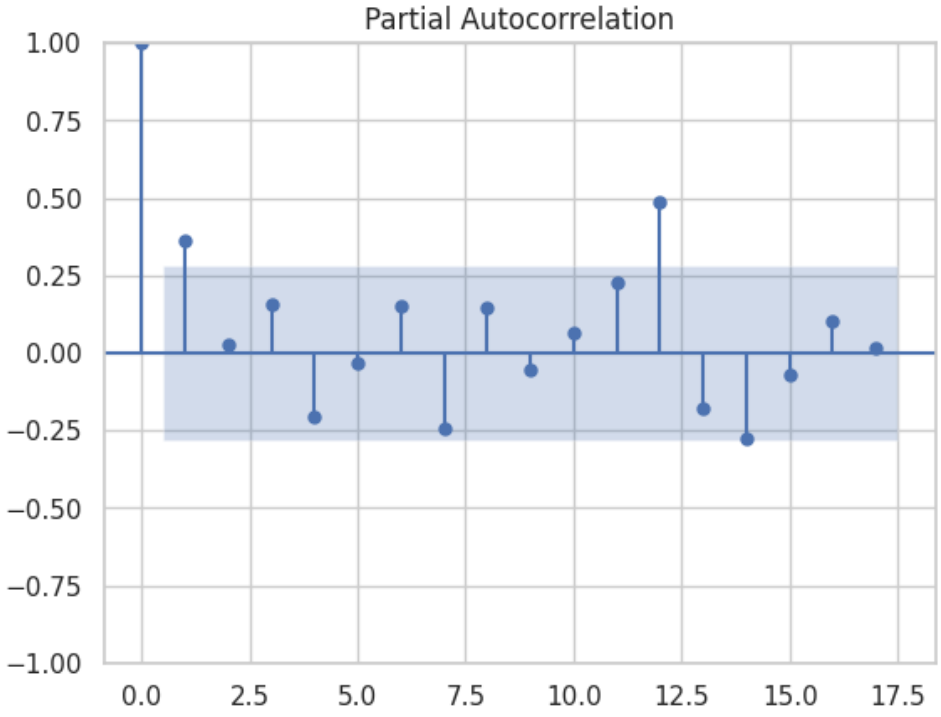

Figure 8: PACF sales

The ACF and PACF plots show strong short-term autocorrelation, supporting the use of an ARIMA(1,1,1) model. The model output indicates the MA(1) term is significant, while AR(1) is not. Diagnostic tests confirm the residuals are white noise (no significant autocorrelation, normal distribution, constant variance), suggesting the model is appropriate for forecasting.

4.3 ARIMA model summary

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	48
Model:	ARIMA(1, 1, 1)	Log Likelihood	-539.895
Date:	Thu, 26 Jun 2025	AIC	1085.789
Time:	11:09:05	BIC	1091.340
Sample:	01-31-2014 - 12-31-2017	HQIC	1087.878
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2278	0.262	0.868	0.386	-0.287	0.742
ma.L1	-0.8396	0.140	-5.979	0.000	-1.115	-0.564
sigma2	5.976e+08	6.93e-11	8.63e+18	0.000	5.98e+08	5.98e+08

Ljung-Box (L1) (Q):	0.28	Jarque-Bera (JB):	1.06
Prob(Q):	0.60	Prob(JB):	0.59
Heteroskedasticity (H):	1.02	Skew:	0.30
Prob(H) (two-sided):	0.96	Kurtosis:	2.58

Figure 9: Model summary

The ACF and PACF plots indicate significant short-term autocorrelation, supporting an ARIMA(1,1,1) structure. In the fitted model, the moving average term (MA1) is statistically significant ($p < 0.01$), while the autoregressive term (AR1) is not ($p = 0.39$). Diagnostic checks show the model residuals are uncorrelated (Ljung-Box $p = 0.60$), normally distributed (Jarque-Bera $p = 0.59$), and homoscedastic (H test $p = 0.96$). Together, these results suggest the model is statistically adequate and well-suited for forecasting future sales.

The ARIMA(1,1,1) model fits the data well. The MA(1) component is significant, while AR(1) is not. Residuals pass key statistical checks, indicating an adequate fit.

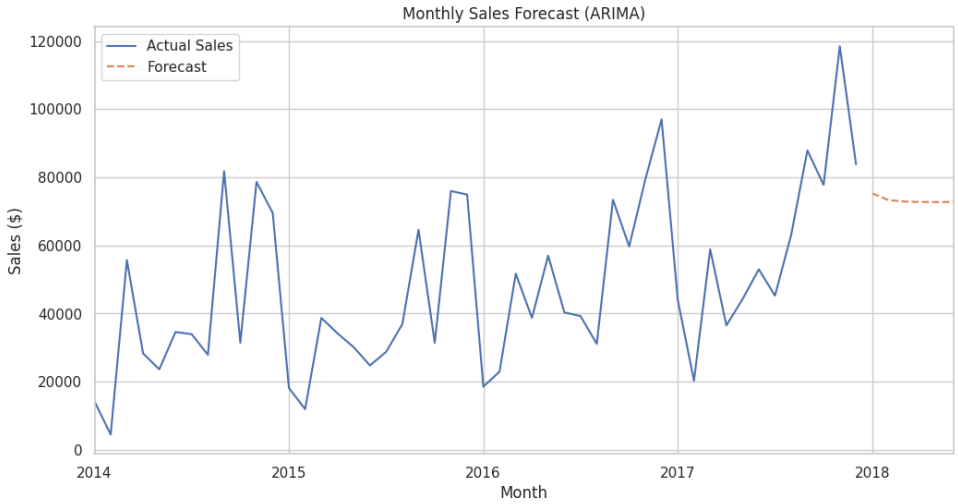

Figure 10: Monthly Sales Forecast using ARIMA

Interpretation: The ARIMA model captures the overall sales trend and gives reasonable short-term forecasts, but may not fully capture seasonality.

4.4 Prophet Model

Prophet, developed by Facebook, is a robust machine learning model for time series forecasting. It can automatically detect trends and seasonal effects. Prophet is an additive model for time series forecasting:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

where:

- $g(t)$: trend function modeling non-periodic changes
- $s(t)$: seasonality
- $h(t)$: effects of holidays
- ε_t: error term

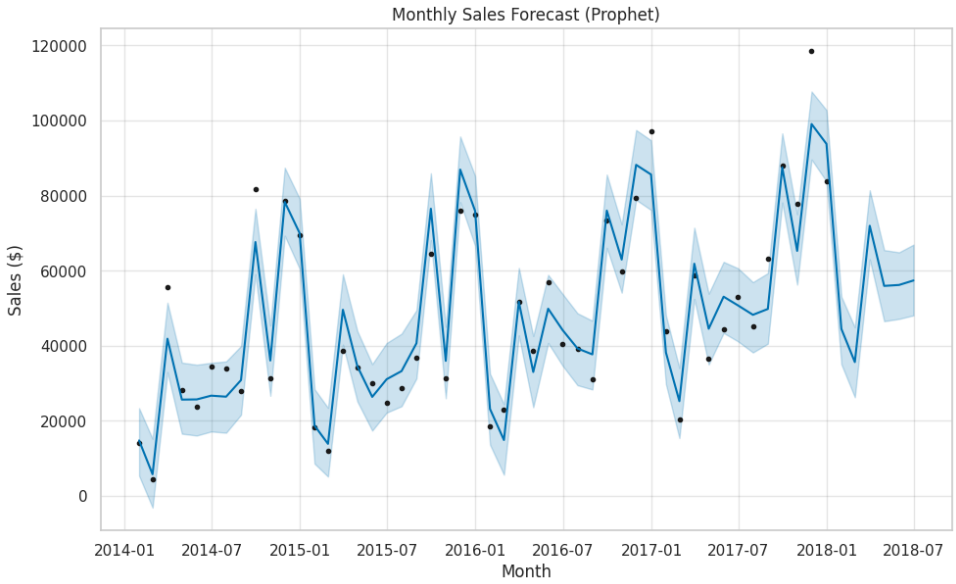

Figure 11: Monthly Sales Forecast using Prophet

This plot shows the Prophet model's monthly sales forecast (solid blue line), with the light blue shaded region representing the model's uncertainty interval (confidence range for future values). The black dots are actual historical sales. The Prophet model captures both the strong upward sales trend and clear seasonality—notice the recurring spikes and dips that align with previous years' patterns. The model predicts that sales will continue to grow, but warns that significant monthly variation (seasonal peaks and troughs) will persist. The width of the shaded area grows slightly in the forecasted period, indicating increased uncertainty further into the future. This information is critical for decision-makers, who must plan for both the expected trend and the possible range of outcomes.

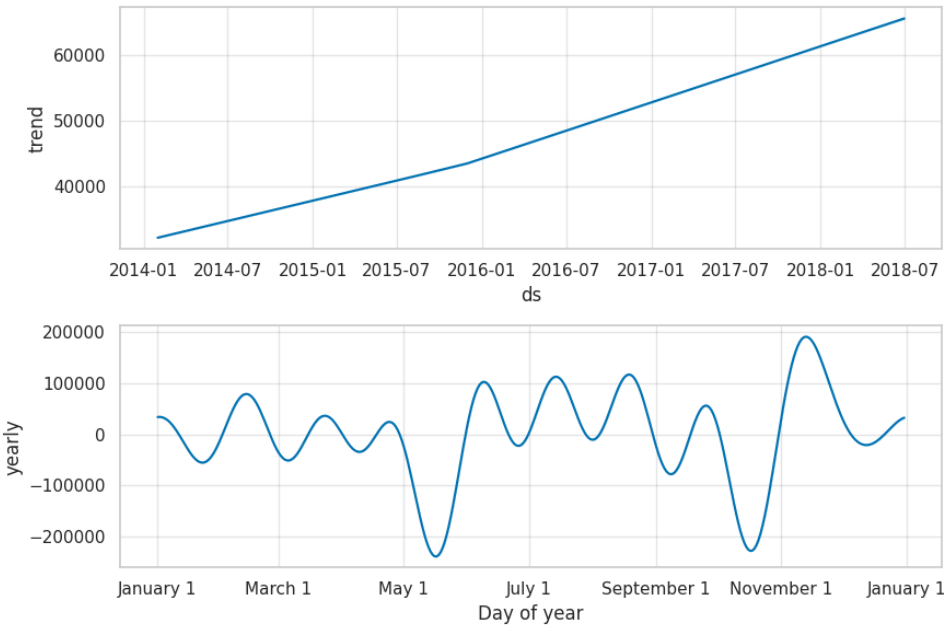

Figure 12: Prophet Components: Trend and Seasonality

The Prophet components plot separates the sales forecast into trend and seasonality. The trend component reveals a consistent upward trajectory in sales over time, while the seasonality component shows strong, repeating annual patterns—highlighting months with expected sales peaks and troughs. These insights help the business plan inventory and marketing to match demand cycles.
The y-axis in the yearly seasonality plot reflects the typical deviation from average sales due to seasonal effects. Positive values indicate above-average months (seasonal peaks), while negative values indicate below-average months (seasonal lows). This allows management to anticipate and plan for predictable fluctuations throughout the year

Interpretation: Prophet provides an interpretable decomposition of the forecast, with a clear upward trend and seasonal cycles, aiding business planning and inventory management.

5 Business Insights and Recommendations

- Sales show strong year-over-year growth, indicating a healthy business trajectory.
- Seasonality is significant; marketing and inventory planning should be adjusted for peak months.
- Technology is the highest-performing category; consider expanding related product lines.
- Underperforming regions (Central, South) may benefit from targeted strategies.
- High discounts often reduce profit; review and optimize discounting policies.

6 Conclusion

This analysis demonstrates the value of combining EDA and advanced forecasting models for actionable retail insights. Both statistical (ARIMA) and machine learning (Prophet) approaches enhance planning and strategy in a competitive environment.

Appendix: Code